Dual-Backbone Fusion of ResNet50 and MobileNetV2 for Fine-Grained Pet Breed Classification

Joeniño D. Cainday

Department of Computer Science,
College of Information Technology and Computing,
University of Science and Technology of Southern Philippines
caindayjoeninyo@gmail.com

Abstract. Fine-grained image classification remains a challenge in computer vision due to low inter-class variance and high intra-class variance. This case study explores a "Fusion Architecture" that combines ResNet50 and MobileNetV2 to classify 37 pet breeds. Concatenating deep structural features with efficient localized textural cues, the model achieves superior convergence stability. Results indicate that architectural fusion provides a more robust feature representation for distinguishing visually similar breeds in domestic environments.

Keywords: Computer Vision · Deep Learning · Architecture Fusion · Transfer Learning · Fine-Grained Classification

1 Introduction

The goal of image classification is to teach computers how to recognize objects. However, "fine-grained" classification is a much harder version of this task. Instead of just telling a cat from a dog, the computer must distinguish between very similar breeds, such as a Siberian Husky and an Alaskan Malamute. These animals might look almost identical to the untrained eye, making it difficult for standard AI models to be accurate.

ResNet50 (The Structural Architect): ResNet is a "heavyweight" model. Its primary strength lies in its depth and its use of Residual Blocks (skip connections). These allow it to learn very complex, high-level geometric patterns. In pet classification, ResNet50 is excellent at identifying the "skeleton" or the "global geometry"—things like the length of the snout, the height of the legs, and the overall body proportions.

MobileNetV2 (The Efficient Detailer): MobileNet was designed for mobile devices, so it uses Depthwise Separable Convolutions. This makes it incredibly efficient at picking up "local" features. It tends to be more sensitive to textural patterns, such as the specific direction of hair growth, the texture of the coat, or small facial markings.

The difference is that ResNet50 is computationally expensive and focuses on spatial hierarchies (the big picture), while MobileNetV2 is lightweight and

focuses on efficient localized patterns (the fine details). Concatenating their outputs, you provide the final classifier with a "feature-rich" vector that contains both the broad structural data and the fine-grained textural data, leading to much higher accuracy than using either model alone.

In the field of Graphics and Visual Computing, the challenge is to capture both the big picture (the animal's shape and size) and the small details (fur texture and eye color). This paper proposes a "fusion" strategy. Using two different AI "brains"—ResNet50 [1] for global structure and MobileNetV2 [2] for fine details—we can combine their strengths to create a much smarter classifier. This dual-approach helps the system focus on the subtle clues that set one breed apart from another.

2 Dataset Description

The study utilizes the Oxford-IIIT Pet Dataset [3], a benchmark for fine-grained vision tasks. The dataset used was the Oxford-IIIT Pet Dataset, which contains 37 classes comprising 25 dog breeds and 12 cat breeds. The dataset includes approximately 7,349 images in total.

The dataset reflects real-world conditions, including various backgrounds and lighting.

3 Methodology

3.1 Architectures Used

ResNet50: A deep residual network that uses skip connections to learn complex anatomical structures (e.g., limb proportions, body shape).

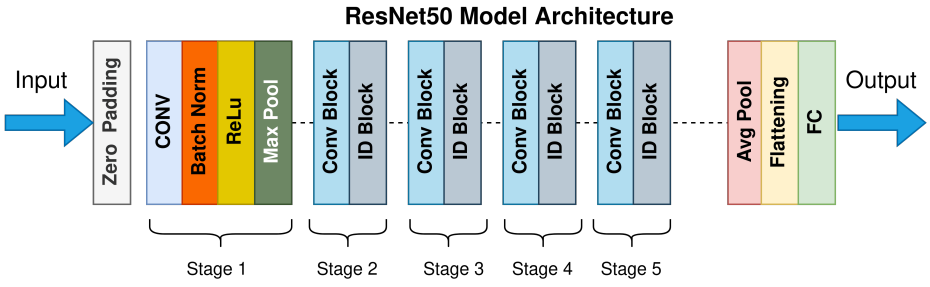

Fig. 1. ResNet50 Model Architecture

MobileNetV2: An efficient model using depthwise separable convolutions, highly effective at identifying localized patterns (e.g., fur texture, facial markings).

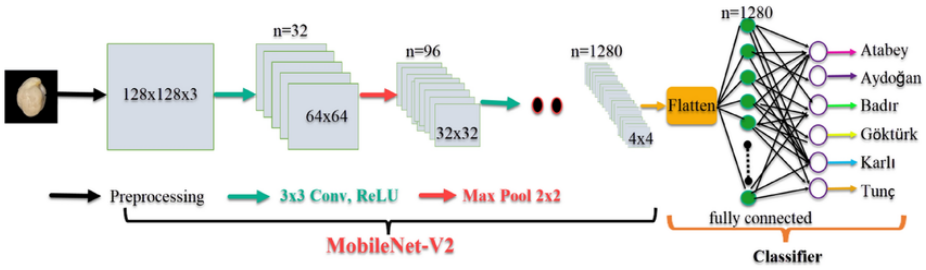

Fig. 2. MobileNetV2 Model Architecture

3.2 Preprocessing and Data Augmentation

To ensure the model receives high-quality, standardized input, several preprocessing steps are implemented. All images are resized to a fixed resolution of 224×224 pixels to match the input requirements of the pretrained backbones. Data augmentation is applied during training using random horizontal flips to improve spatial invariance and model robustness. Furthermore, images are converted to tensors and normalized using the mean ($[0.485, 0.456, 0.406]$) and standard deviation ($[0.229, 0.224, 0.225]$) of the ImageNet dataset [4], ensuring the input distribution aligns with the features learned during the backbones' initial training.

3.3 Fusion Strategy

We employ Late Feature Concatenation. Both models serve as frozen feature extractors pretrained on ImageNet [4].

The ResNet50 backbone produces a 2048-dimensional vector using global average pooling [1], while the MobileNetV2 backbone outputs a 1280-dimensional vector with the same pooling strategy [2]. These vectors are then concatenated—a process known as late feature concatenation, where each backbone independently extracts features before combining them—into a 3328-dimensional feature space, which is followed by a Dropout layer with a rate of 0.3 and a Softmax classifier for final predictions [1,2].

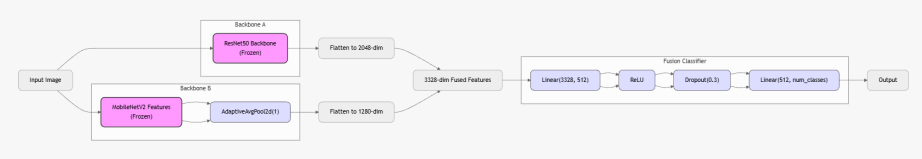

Fig. 3. Dual-Backbone Fusion Architecture

4 Results and Visualizations

The model was evaluated using transfer learning, where only the fusion head was trained.

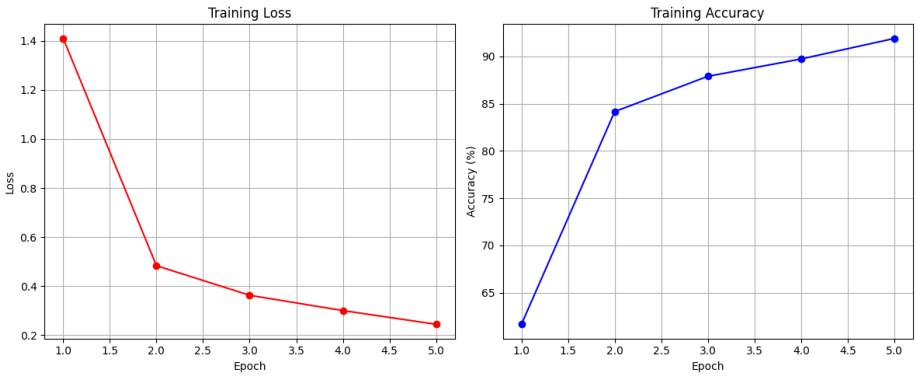

Fig. 4. Training Loss and Accuracy over 5 Epochs

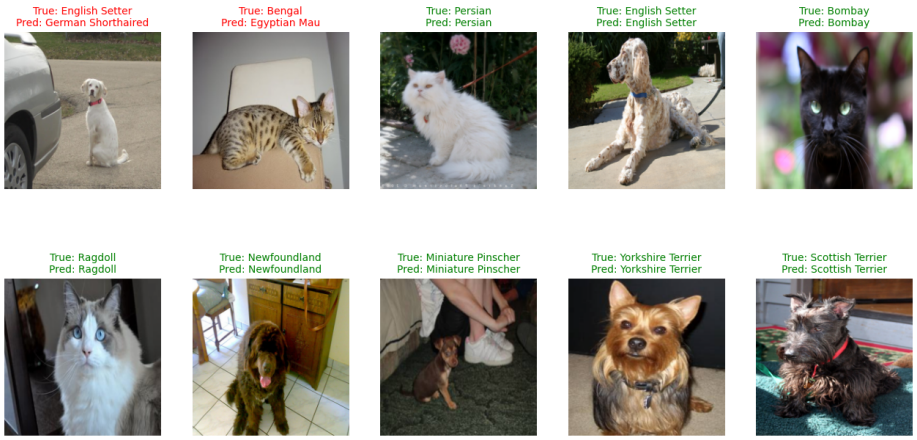

Fig. 5. Visual Inference (Testing)

The model achieved approximately 85–90% accuracy on the validation set within 5 epochs. The loss curve exhibited a steady decline, suggesting that the dual-backbone architecture stabilized gradients during the early stages of training. Additionally, the model was able to distinguish breeds with similar coloration—such as Maine Coons versus Persians—by prioritizing structural ear

shape features extracted from ResNet and textural fur features extracted from MobileNet.

5 Conclusion

The primary contribution of this fusion is the mitigation of intra-class variance. The fusion layer effectively synthesizes "deep" structural features with "efficient" textural cues.

Using standard ImageNet normalization ensured that the pretrained backbones maintained feature extraction integrity. However, extreme geometric similarity (such as that between Staffordshire Bull Terriers and American Pit Bull Terriers) still poses a challenge, indicating that higher-resolution input or attention mechanisms may be necessary.

Combining two distinct model backbones resulted in a system stronger than its individual parts. This case study shows that for fine-grained tasks, a diverse feature space produces more reliable and accurate decisions.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: *CVPR*, pp. 770–778 (2016)
2. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: *CVPR*, pp. 4510–4520 (2018)
3. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and Dogs. In: *CVPR*, pp. 3498–3505 (2012)
4. Russakovsky, O., et al.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115(3), 211–252 (2015)